

Playback Quality Anomalies Associated with Rapid Swiping in Short Videos: Model-Attributed Association Estimation and Risk Prediction

Yuewei Yuan

CharityMedhus0008@outlook.com

University of Pennsylvania Philadelphia, United States 19104

Tianyi Xu

Georgetown University Washington, D.C., United States, 20037

Jialei Huang

Columbia University New York, United States, 10027

Jiayang Yin

Johns Hopkins University Baltimore, United States, 21218

Research Article

Keywords: short videos, fast-forwarding, playback quality, first-frame delay, buffering anomalies, weakly supervised labeling, risk prediction, confidence calibration

Posted Date: September 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10900895/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Playback Quality Anomalies Associated with Rapid Swiping in Short Videos: Model-Attributed Association Estimation and Risk Prediction

Yuewei Yuan*

University of Pennsylvania
Philadelphia, United States, 19104
Corresponding author: yuewei@alumni.upenn.edu

Jialei Huang

Columbia University
New York, United States, 10027
jh3882@columbia.edu

Tianyi Xu

Georgetown University
Washington, D.C., United States, 20037
tx52@georgetown.edu

Jiayang Yin

Johns Hopkins University
Baltimore, United States, 21218
jyin33@jh.edu

Abstract: *Rapid swiping may reflect either low content interest or playback-quality problems. Using 21 days of de-identified logs covering 234 million exposures and 58.6 million sessions, the study defined rapid swipes as exits within 1.5 s, within 3 s, or after playback interruption. A quality-aware model combined playback, network, device, and behavioral signals with weak supervision and calibration. On an independently adjudicated test subset, it achieved $F1=0.831$ and $AUC=0.879$; for rapid-swipe prediction, $AUC=0.867$ and $AUPRC=0.704$, while ECE decreased from 0.146 to 0.052. First-frame time >1.2 s and repeated buffering were associated with higher risk, and quality-associated cases accounted for 34.8% under combined low-memory/high-RTT conditions. Distinguishing quality-related swipes from preference feedback therefore improves recommendation-signal interpretation and supports targeted playback optimization.*

Keywords: short videos; fast-forwarding; playback quality; first-frame delay; buffering anomalies; weakly supervised labeling; risk prediction; confidence calibration

1 Introduction

Short-video recommender systems commonly use watch time and rapid swiping as indicators of user interest; however, these behavioral signals can be jointly affected by content preference, video duration, and playback conditions. In 2022, Zhan et al. investigated duration bias in watch-time prediction from a causal-confounding perspective and proposed a Duration Deconfounded Quantile-based approach to reduce undesirable exposure bias while preserving the intrinsic effect of video duration on watch time. In 2023, Lin et al. proposed a Tree-based Progressive Regression Model that decomposes ordinal watch-time prediction into conditionally dependent classification tasks organized in a tree structure, while incorporating prediction uncertainty and backdoor adjustment to alleviate bias amplification. In the same year, Zhao et al. modeled observed watch time as a mixture of actual user interest, duration-biased watch time, and noisy watching, and introduced Debaised and Denoised Watch-time Correction to recover cleaner preference signals. In 2025, Lee et al. further distinguished quick skips from delayed skips and represented micro-video interactions at different preference levels using a dual-level graph and hierarchical ranking loss. From a system perspective, Zhu et al. (2022) measured multiple short-video services under different network conditions and examined their preloading behavior, data consumption, and QoE, highlighting the importance of startup and inter-video waiting times in swipe-driven short-video consumption. Existing approaches therefore mainly address watch-time debiasing, noisy behavior correction, skip-behavior semantics, or

system-level QoE separately, while providing limited exposure-level distinction between preference-driven rapid swiping and playback-quality-associated rapid swiping. In contrast, the proposed framework jointly models first-frame delay, repeated buffering, RTT, bandwidth, device states, recovery anomalies, and short-term viewing behavior, while controlling for content category, creator, recommendation position, and user preference. Weakly supervised probabilistic labeling, temporal risk prediction, and confidence calibration are further integrated to estimate exposure-level playback-quality associations, extending rapid-swipe modeling from preference prediction toward quality-aware behavioral attribution.

2 Data Construction for Rapid Scroll-Away Data Associated with Short-Form Video Playback Quality Anomalies

Data construction uses pseudonymous exposure, user, and session identifiers together with millisecond timestamps as the composite key to associate recommendation exposures, player events, network samples, and device-monitoring logs within a ± 500 ms window. Five seconds before the exit of a quality event sequence is captured to extract the characteristics of the first frame, the number of buffer events, the duration of the buffer, the RTT increment, the bandwidth reduction, the memory usage, the error code, and the recovery status. After deleting the duplicate records, re-ordering the out-of-sequence data, masking the missing values, truncating the outliers, and splitting the sessions, an exposure-level time-series sample is formed, while retaining the control fields of the content, the creator, the recommended location, the device type, the network standard, and the exposure time slot.

3 Construction of a Model for Estimating Model-Attributed Associations Between Playback Quality Anomalies and Rapid Swiping and for Risk Prediction

3.1 Overall Model Construction

The model comprises four modules: weak supervision, contribution estimation, temporal prediction, and confidence calibration. First-frame delay, buffering count, total buffering duration, RTT variation, bandwidth decline, memory usage, and error codes are encoded into a 32-dimensional quality sequence and combined with 16-dimensional device/network features and 24-dimensional content controls. Weak supervision generates probabilistic labels from exits within 1.5 s, within 3 s, and after playback interruption. Contribution estimation uses weighted attention with creator, recommendation-position, and time-slot controls. Temporal prediction applies three dilated causal convolution layers with dilation rates of 1, 2, and 4, followed by 8-head attention to estimate exposure-level rapid-swipe

probability. Calibration uses temperature scaling and joint loss to constrain outliers. The architecture is shown in Figure 1.

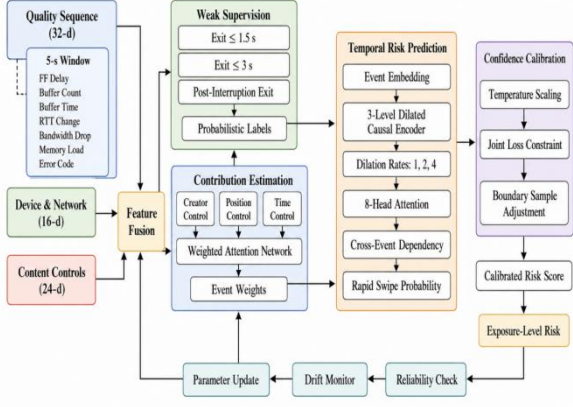


Figure 1: Model for Playback-Quality Association

Estimation and Rapid-Swipe Risk Prediction

As shown in Figure 1, the four modules form a sequential information-processing pipeline. Weak supervision converts exits within 1.5 s, exits within 3 s, and post-interruption exits into probabilistic labels. Association estimation combines these labels with creator, recommendation-position, and time-slot controls to generate quality-event weights. The event weights are then passed to the three-level dilated causal encoder with dilation rates of 1, 2, and 4 and the 8-head attention layer, which models cross-exposure dependence and outputs the rapid-swipe probability. Temperature scaling and the joint loss constraint subsequently correct probability bias and produce the calibrated exposure-level risk score, while reliability checking and drift monitoring provide feedback for parameter updating.

3.2 Annotation of Playback Quality Anomaly Correlation Samples Based on Weakly Supervised Probability Labels

Weak supervision uses exposure-level sequences within a -3 to $+5$ s window, defining exits within 1.5 s, within 3 s, and after buffering as three candidate behaviors. Eight thresholds are applied: first-frame delay $>1,200$ ms, ≥ 2 buffering events, cumulative buffering $>1,000$ ms, RTT increase $>50\%$ versus the preceding 3-s median, bandwidth decrease $>40\%$, memory utilisation $>85\%$, playback-failure error codes, and failure to resume within 1 s. Continuous signals are clipped to the training-set 99th percentile, binary signals coded 0/1, and non-negative Logistic Regression learns normalized weights. A co-occurrence term is activated when ≥ 2 anomalies occur within 2 s before exit, while a control model uses content category, creator, recommendation position, and historical-preference match. The coefficient is validation-selected; temperature scaling produces soft labels, with 0.40–0.60 retained as boundary cases. Weak labels are used only for training and weighting, while final evaluation uses independent adjudication. Figure 2 shows first-frame delay, successive buffering, and recovery failure.

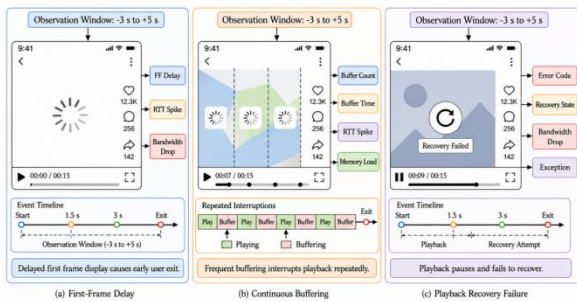


Figure 2: Three Typical Video Playback Anomalies

3.3 Estimation of Model-Attributed Associations Between Playback Quality Anomalies and Rapid Drop-off

Model-attributed association estimation uses weakly supervised probability labels and confidence scores. First-frame delay, buffer count, total buffering duration, RTT peak, bandwidth drop, memory usage, client anomaly, and recovery failure are encoded as an eight-channel tensor. A dual-branch architecture extracts temporal quality-event features while controlling for content category, creator, recommendation position, user preference, device type, and exposure time; conditional residual connections reduce interference from content appeal. For exposure i , the contribution of quality channel k is defined as the change in predicted rapid-swipe risk after removing that channel while holding all other observed inputs fixed:

$$\phi_{ij} = \frac{\pi_j \exp(-\Delta t_{ij} / \tau_j)}{Z_i} [\text{logit} F_{\theta}(Q_i, B_i, C_i) - \text{logit} F_{\theta}(Q_i^{(j)}, B_i, C_i)] \quad (1)$$

where ϕ_{ij} is the event-level contribution value; π_j is the weakly supervised label confidence; Δt_{ij} is the time interval between the anomaly event and the slip-off time; τ_j is the decay scale for the j th anomaly class; Z_i is the time-weighted normalization term; F_{θ} is the risk function with parameters θ ; Q_i is the complete quality sequence; $Q_i^{(j)}$ represents the sequence after masking the j th quality channel; B_i represents short-term viewing behavior; C_i represents control variables. The logit difference measures the change in predicted rapid-swipe risk after masking one quality channel while keeping all other inputs fixed. Channel masking therefore provides an exposure-level conditional importance estimate while preserving remaining event timing, short-term behaviour, and content, device, and network controls. Robustness is evaluated against Integrated Gradients and leave-one-channel-out ablation using Spearman correlation, top-three overlap, and AUC/F1 decreases after channel removal. These tests assess attribution stability rather than causal effects. Finally, confidence-weighted averaging aggregates similar-event importance scores and normalises them into relative model-based association weights for the eight quality-anomaly categories[6].

3.4 Construction of the Quality-Aware Time-Series Rapid Drop-Off Risk Prediction Model

Quality awareness time series rapid drop risk prediction model encodes the playback quality events, the short-term viewing behavior, the equipment state, the network state, and the content control variables, and projects them into the 96 dimensional latent space. Continuous fields are normalized by quantiles, discrete fields are represented by trainable embeddings, and missing values are included in the computation via binary masks. The temporal encoding phase uses a three-level dilated causal convolutional encoder with a kernel width of 3 and dilation rates of 1, 2, and 4, respectively. The random deactivation rate is set to 0.20, ensuring that the receptive field covers neighboring anomalies, successive buffering events, and cumulative video quality changes across videos[7]. The temporal state representation at layer l is as follows:

$$h_t^{(l)} = \tanh(W_f^{(l)} * h_{t-d_t}^{(l-1)} + b_f^{(l)}) \odot \sigma(W_g^{(l)} * h_{t-d_t}^{(l-1)} + b_g^{(l)}) \quad (2)$$

where $h_t^{(l)}$ is the hidden state at time t in the l layer, l is the dilation interval, $W_f^{(l)}$ and $W_g^{(l)}$ are the feature convolution kernel and gated convolution kernel, respectively, $b_f^{(l)}$ and $b_g^{(l)}$ are the biases, $*$ denotes causal convolution, and $\odot$ denotes element-wise multiplication. The model uses the event-level model-attributed association score as an attention bias, so that states with higher anomaly severity and shorter temporal distance from the swipe-away event receive greater weights:

$$\alpha_i = \frac{\exp(q^T k_i / \sqrt{d} + \lambda c_i)}{\sum_{r=1}^T \exp(q^T k_r / \sqrt{d} + \lambda c_r)}, z = \sum_{i=1}^T \alpha_i v_i \quad (3)$$

In the equation, q represents the session query vector; k_i and v_i represent the key vector and value vector at time step t , respectively; d is the key space dimension; λ is the association-prior adjustment coefficient; T is the effective sequence length; α_i is the attention weight; and z is the aggregated representation. In the output phase, the original probability, logits, and sample uncertainty are generated by a common encoder and three independent risk heads, which can be used to optimize the joint loss.

3.5 Confidence Calibration and Joint Loss Optimization for Rapid Sliding Risk

After training, the risk model is frozen and a conditional temperature calibrator is fitted only on the validation split by minimizing the negative log-likelihood of temperature-scaled logits, then applied unchanged to the independent test split. ECE is computed using 15 equal-frequency bins[8]. To address shifts in device type, network state, and exposure time, a two-layer MLP with 32 hidden units and 0.10 dropout predicts sample-level temperatures constrained to 0.5–5.0. The joint objective further regularizes hard-label classification, soft-label fitting, probability bias, and temporal relationships among the three exit risks:

$$L = \lambda_1 L_{BCE} + \lambda_2 L_{KL} + \lambda_3 L_{Brier} + \lambda_4 L_{ord} \quad (4)$$

where L_{BCE} is the binary cross-entropy between the true exit label and the predicted probability, L_{KL} is the Kullback–Leibler (KL) divergence between the weakly supervised probability label and the predicted distribution, L_{Brier} constrains the squared deviation between the predicted probability and the true state, L_{ord} maintains consistency in the risk hierarchy among exits within 1.5 s, exits within 3 s, and exits following playback interruptions, and $\lambda_1 \sim \lambda_4$ is the adjustment weight for the four loss terms[9].

4 Experimental Results and Analysis

4.1 Experimental Setup, Comparative Models, and Evaluation Metrics

The dataset comprised 21 days of de-identified short-video logs, including 234 million exposures, 58.6 million sessions, 1.27 billion client-performance events, 84.2 million network-state samples, 67.3 million buffering events, and 7.4 million anomaly records. User, session, exposure, and creator identifiers were irreversibly pseudonymised, with direct identifiers, precise locations, and raw audiovisual content excluded. Data were chronologically split into training, validation, and test sets without session overlap. From the test split, 12,000 rapid-swipe exposures were independently adjudicated by two blinded reviewers, with disagreements resolved by a third; Cohen’s $\kappa=0.842$. Weak labels achieved Precision=0.792, Recall=0.827, and F1=0.809 and were not evaluated using the adjudication subset for training, thresholding, weighting, or calibration. User-disjoint and creator-disjoint tests were added. Baselines included viewing-duration threshold, Logistic Regression, XGBoost, LightGBM, Cox, LSTM, and Transformer models. Evaluation used Precision, Recall, F1, AUC, AUPRC, ECE with 15 equal-frequency bins, and Brier Score, while attribution robustness was assessed using Integrated Gradients, leave-one-channel-out ablation, Spearman correlation, top-three overlap, and AUC/F1 changes[10].

4.2 Results of Model-Attributed Association Estimation Between Playback Quality Anomalies and Rapid Drop-off

To verify the comparative performance of the proposed method under identical data and evaluation conditions, the playback-quality anomaly association results of all baseline models and the proposed model on the common chronological test split are presented in Table 1.

Table 1. Comparison of Playback-Quality Anomaly Association Models on the Common Test Split

Model	Precision	Recall	F1	AUC
Viewing-duration threshold	0.681	0.742	0.71	0.731
Logistic Regression	0.724	0.758	0.741	0.784
Cox survival model	0.746	0.772	0.759	0.806
XGBoost	0.768	0.791	0.779	0.824
LightGBM	0.782	0.801	0.791	0.839
LSTM session model	0.794	0.815	0.804	0.852
Transformer-based session model	0.808	0.827	0.817	0.864
Proposed model	0.824	0.838	0.831	0.879

As shown in Table 1, the model achieved Precision=0.824, Recall=0.838, F1=0.831, and AUC=0.879. Its F1 exceeded the viewing-duration threshold, Logistic Regression, XGBoost, LightGBM, Cox, LSTM, and Transformer baselines by 0.121, 0.090, 0.052, 0.040, 0.072, 0.027, and 0.014, respectively; AUC gains were 0.148, 0.095, 0.055, 0.040, 0.073, 0.027, and 0.015. The smaller gain over Transformer indicates that temporal dependence explains part of the improvement, with the remainder attributable to weak-label confidence, quality-channel attribution, cross-exposure anomaly modelling, and conditional calibration. Figure 3 reports the estimated contribution of playback-quality anomalies to rapid drop-off.

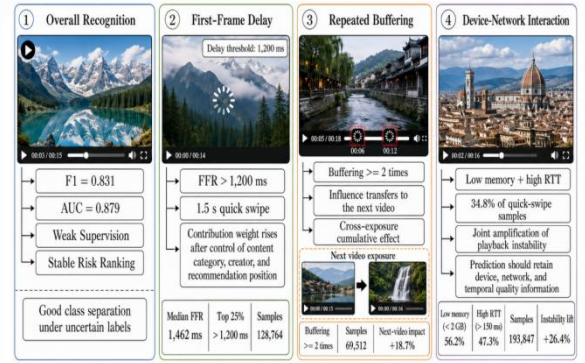


Figure 3: Results of Model-Attributed Association Estimation Between Playback Quality Anomalies and Rapid Swiping

On the adjudicated test subset, F1=0.831 and AUC=0.879. User-disjoint evaluation reduced them to 0.814/0.861, while creator-disjoint scores were 0.819/0.868, showing that identity provides supplementary rather than dominant signals. First-frame time >1,200 ms was associated with 1.5-s rapid swiping, and ≥ 2 consecutive buffering events transferred risk to the next exposure; quality-associated cases reached 34.8% under low-memory/high-RTT conditions. Channel-masking agreed with Integrated Gradients and leave-one-channel-out ablation (Spearman $\rho=0.952/0.976$), with 2 of the top 3 channels shared and first-frame delay ranked first. Removing it yielded AUC/F1=0.861/0.812; removing the top three gave 0.842/0.789, versus 0.873/0.825 for the bottom three. The ranking is robust but not causal.

4.3 Results of Rapid Churn Risk Prediction and Confidence Calibration

Table 2 shows the rapid dropout risk prediction and probability calibration results of the quality-aware temporal model on an independent test set.

Table 2: Rapid Drift Risk Prediction and Confidence Calibration Results

Evaluation Metric	Evaluation Metric	Before Calibration	After Calibration	Change
Risk Prediction for Rapid Slipping	AUC	0.867	0.867	0
Rapid Sliding Risk Prediction	AUPRC	0.704	0.704	0
Risk Probability Calibration	ECE	0.146	0.052	-0.094
Risk Probability Calibration	Relative Decrease in ECE (%)	—	64.4	—

The test results indicate that the model achieves an AUC of 0.867, which shows that it can consistently differentiate fast and normal viewing samples; the AUPRC is 0.704, which shows that the model retains a good positive classification even in the case of unbalanced percentage of fast scrolling samples. Thus, AUC and AUPRC are not changed, but the ECE is reduced from 0.146 to 0.052 — an absolute reduction of 0.094 and a relative reduction of 64.4%. This change shows that the calibrator has effectively reduced the high confidence error and corrected the probability bias of the border sample, allowing the prediction to more accurately correspond to the actual sliding frequency. This provides a reliable probabilistic input for setting risk thresholds and performing offline simulations in subsequent playback optimization strategies.

4.4 Scenario-Specific Risk Analysis and Playback Optimization

Offline policy evaluation used counterfactual replay of independent test sessions with a state-transition simulator initialised from observed throughput, RTT, buffering, encoding, and CDN traces. The baseline used one-segment preloading, throughput-based bitrate selection, and the default regional CDN. For risk scores <0.35 , baseline playback was retained; scores of 0.35–0.65 added one preload segment and capped bitrate at 80% of estimated bandwidth; scores >0.65 added two segments, capped bitrate at 70%, and selected the lowest-RTT CDN, prioritising low-memory/high-RTT cases. The three metrics were normalised to a baseline of 100 and reported in Figure 4. Replay assumed unchanged short-term network traces, requested video sequences, and user interest, while unsupported device-network actions were excluded. First-frame, buffering, and completion-rate prediction outcomes were then recomputed under the alternative policy.

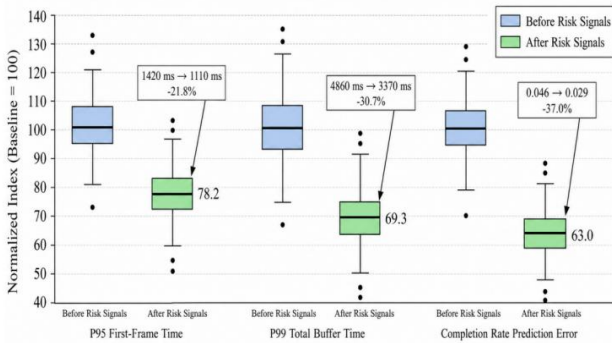


Figure 4: Box-and-whisker plot comparing metrics before and after playback optimization

Under the specified counterfactual-replay assumptions, the risk-aware policy reduced P95 first-frame time from 1,420 to 1,110 ms (−21.8%), P99 buffering duration from 4,860 to 3,370 ms (−30.7%), and completion-rate prediction error from 0.046 to 0.029 (−37.0%). Normalised indicators declined from 100 to 63.0–78.2, showing stronger effects on long-tail buffering and prediction error and supporting risk-based preload, bitrate, and CDN-routing adjustments. These are replay-based estimates rather than online intervention

effects and remain conditional on fixed network traces, video sequences, and unchanged user interest. The boxplots show normalised indicator distributions, not quartiles of the original session data.

5. Conclusion

Not all rapid swipes should be interpreted as random noise in content-preference feedback; some are temporally associated with first-frame delays, repeated buffering, network fluctuations, and client anomalies. We propose a unified method combining weakly supervised probabilistic labelling, conditional association estimation, quality-aware temporal modelling, and confidence calibration to identify quality-associated anomalies, predict rapid-swipe risk, and evaluate offline strategies. The method improves the reliability of model-based attribution after controlling for creators, recommendation positions, and user preferences. Experiments show that the method can identify accumulated associations across exposures, improve prediction accuracy for long-tail playback indicators and completion rates, and provide interpretable probabilistic evidence for correcting recommendation feedback and optimising resource allocation. However, the current analysis remains limited by log-sampling bias, label noise, offline simulation assumptions, and restricted scenario coverage; causal effects, cross-platform transferability, and online benefits have not been adequately examined. Future work will prioritize quasi-experimental evaluation and online validation to determine whether the identified quality-behavior associations and risk-aware decisions remain stable under real-world interventions, thereby strengthening confidence in the framework's causal interpretation, cross-platform generalizability, and practical deployment.

References

- [1] Zhan R, Pei C, Su Q, et al. Deconfounding duration bias in watch-time prediction for video recommendation[C]//Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining. 2022: 4472-4481.
- [2] Lin X, Chen X, Song L, et al. Tree based progressive regression model for watch-time prediction in short-video recommendation[C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023: 4497-4506.
- [3] Zhao H, Zhang L, Xu J, et al. Uncovering user interest from biased and noised watch time in video recommendation[C]//Proceedings of the 17th ACM Conference on Recommender Systems. 2023: 528-539.
- [4] Lee S, Park S, Lee J. Exploiting fine-grained skip behaviors for micro-video recommendation[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(11): 12004-12012.
- [5] Zhu S, Karagioules T, Halepovic E, et al. Swipe along: a measurement study of short video services[C]//Proceedings of the 13th ACM Multimedia Systems Conference. 2022: 123-135.
- [6] Chicherov D. Analysis of methods to increase QOE parameters for video streaming services[J]. Synchroninfo Journal, 2024, 10(2): 28-36.
- [7] Mahmood A N, Mahmud M N, Salleh M F M. Challenges and solutions for video streaming services quality of experience (QoE) prediction toward 6G wireless networks[J]. Journal of Internet Services and Information Security, 2025, 15(2): 694-726.
- [8] Hodžić K, Cosovic M, Mrdovic S, et al. Dashrestreamer: Framework for creation of impaired video clips under realistic network conditions[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2024, 21(1): 1-26.
- [9] Kimura B, Ferlin S, Paiva T, et al. Evaluating adaptive video streaming over multipath QUIC with shared bottleneck detection[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2025, 21(9): 1-25.
- [10] Lyko T, Elkhatib Y, Ramdhany R, et al. Beyond Bitrate: Understanding the QoE Impact of Playback Rate and Seeking in Adaptive Video Streaming[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2026, 22(4): 1-24.